Reinforcement Learning with Depreciating Assets

Extended Abstract

Taylor Dohmen
University of Colorado
Boulder, CO, USA
taylor.dohmen@colorado.edu

Ashutosh Trivedi
University of Colorado
Boulder, CO, USA
ashutosh.trivedi@colorado.edu

ABSTRACT

A basic assumption of traditional reinforcement learning is that the value of a reward does not change once it is received by an agent. The present work forgoes this assumption and considers the situation where the value of a reward decays proportionally to the time elapsed since it was obtained. Emphasizing the inflection point occurring at the time of payment, we use the term *asset* to refer to a reward that is currently in the possession of an agent. Adopting this language, we initiate the study of depreciating assets within the framework of infinite-horizon quantitative optimization. In particular, we propose a notion of asset depreciation, inspired by classical exponential discounting, where the value of an asset is scaled by a fixed discount factor at each time step after it is obtained by the agent. We formulate an equational characterization of optimality in this context, establish that optimal values and policies can be computed efficiently, and develop a model-free reinforcement learning approach to obtain optimal policies.

KEYWORDS

Reinforcement Learning; Temporal Discounting; Asset Depreciation

ACM Reference Format:
Taylor Dohmen and Ashutosh Trivedi. 2023. Reinforcement Learning with Depreciating Assets: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Time preference [10, 13] refers to the tendency of rational agents to value potential *desirable outcomes* in proportion to the expected time before such an outcome is realized. In other words, agents prefer to get a future reward sooner rather than later, all else being equal, and similarly, agents prefer to experience negative outcomes later rather than sooner. This phenomenon is typically codified in mathematical models in terms of discounting [17] and has been applied to a diverse array of disciplines concerned with optimization such as economics [11, 14], game theory [9], control theory [15], and reinforcement learning [19]. These models focus on the situation in which an agent moves through a stochastic environment in discrete time by selecting an action to perform at each time step and receiving an immediate reward based on the selected action and environmental state. In particular, we consider exponential discounting, as introduced by Shapley [17], in which the agent carries this process on ad infinitum to generate an infinite sequence of rewards $\langle r_n \rangle_{n=1}^{\infty}$ with the goal of maximizing, with respect to a discount factor $\lambda \in (0, 1)$, the discounted sum $\sum_{n=1}^{\infty} \lambda^{n-1} r_n$. The discount factor is selected as a parameter and quantifies the magnitude of the agent's time preference.

A notable characteristic of the aforementioned discounted optimization framework is an implicit assumption that the utility of a reward remains constant once it is obtained by a learning agent. While this seemingly innocuous supposition simplifies the model and helps to make it amenable to analysis, there are a number of scenarios where such an assumption is not appropriate. Consider, for instance, the most basic and ubiquitous of rewards used to incentivize human behaviors: money. The value of money tends to decay with time according to the rate of inflation, and the consequences of this decay are a topic of wide spread interest and intense study [2, 5, 8, 12]. *Recognizing the fundamental role such decay has in influencing the dynamics of economic systems throughout the world, we consider its implications with respect to optimization and reinforcement learning in Markov decision processes.*

2 ASSET DEPRECIATION

When discussing a situation with decaying reward values, it is useful to distinguish between potential future rewards and actual rewards that have been obtained. As such, we introduce the term *asset* to refer to a reward that has been obtained by an agent at a previous moment in time. Using this terminology, the present work may be described as an inquiry into optimization and learning under the assumption that assets *depreciate*. Depreciation, a term borrowed from the field of finance and accounting [4, 21], describes exactly the phenomenon where the value of a commodity or asset decays with time.

We propose a notion of depreciation that is inspired by traditional discounting and is based on applying the same basic principle of time preference to an agent's history in addition to its future. More precisely, we consider the situation in which an agent's behavior is evaluated with respect to an infinite sequence of cumulative accrued assets, each of which is discounted in proportion to how long ago it was obtained. That is, we propose evaluating the agent in terms of functions on the sequence of assets $\left\langle \sum_{k=1}^{n} r_k \gamma^{n-k} \right\rangle_{n=1}^{\infty}$, where $\gamma \in (0, 1)$ is a discount factor, rather than on the sequence of rewards $\langle r_n \rangle_{n=1}^{\infty}$. To motivate the study of depreciation and argue its naturalness, we examine the following hypothetical case-study.

Example 2.1 (Used Car Dealership). Consider a used car dealership with a business model involving purchasing used cars in locations with favorable regional markets, driving them back to their shop, and selling them for profit in their local market. Suppose that our optimizing agent is an employee of this dealership, tasked with managing capital acquisition. More specifically, this

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

employee's job is to decide the destination from which the next car should be purchased, whenever such a choice arises. The objective of the agent is to maximize the sum of the values of all vehicles in stock at the dealership over a discounted time-horizon for some discount factor $\lambda \in (0,1)$. Note that the discounted time-horizon problem is equivalent to the problem of maximizing expected terminal payoff of the process given a constant probability $(1-\lambda)$ of terminating operations at any point.

It has long been known [1, 22] that cars tend to continually depreciate in value after being sold as new, and so any reasonable model for the value of all vehicles in the inventory should incorporate some notion of asset depreciation. Suppose that another discount factor $\gamma \in (0,1)$ captures the rate at which automobiles lose value per unit of time. Considering γ-depreciated rewards and λ-discounted horizon, the goal of our agent can be defined as a *discounted depreciating optimization problem*. Alternatively, one may seek to optimize the *long run average* (mean payoff) [15] of γ-depreciated rewards.

3 DISCOUNTED DEPRECIATING PAYOFF

Consider a sequence $x = \langle 3,4,5,3,4,5,\ldots\rangle$ of (absolute) rewards accumulated by the agent. In the presence of depreciation, the cumulative asset values at various points in time follow the sequence

$$3,\ (3\gamma{+}4),\ (3\gamma^2{+}4\gamma{+}5),\ (3\gamma^3{+}4\gamma^2{+}5\gamma{+}3),\ (3\gamma^4{+}4\gamma^3{+}5\gamma^2{+}3\gamma{+}4),\ldots$$

The λ-discounted value of this sequence can be computed as follows:

$$\begin{aligned}
&3 + \lambda(3\gamma{+}4) + \lambda^2(3\gamma^2{+}4\gamma{+}5) + \lambda^3(3\gamma^3{+}4\gamma^2{+}5\gamma{+}3) + \\
&\quad \lambda^4(3\gamma^4{+}4\gamma^3{+}5\gamma^2{+}3\gamma+4) + \ldots \\
&= (3{+}3\lambda\gamma{+}3\gamma^2\lambda^2{+}\cdots) + (4\lambda{+}4\lambda^2\gamma{+}4\lambda^3\gamma^2{+}\cdots) + \\
&\quad (5\lambda^2{+}5\lambda^3\gamma{+}\lambda^5\gamma^2{+}\cdots) + (3\lambda^3{+}3\lambda\gamma^4{+}3\gamma^2\lambda^5{+}\cdots) + \cdots \\
&= 3(1{+}\lambda\gamma{+}\gamma^2\lambda^2{+}\cdots) + 4\lambda(1{+}\lambda\gamma{+}\lambda^2\gamma^2{+}\cdots) + \\
&\quad 5\lambda^2(1{+}\lambda\gamma{+}\lambda^2\gamma^2{+}\cdots) + 3\lambda^3(1{+}\lambda\gamma{+}\gamma^2\lambda^2{+}\cdots) + \ldots \\
&= \frac{3+4\lambda+5\lambda^2+3\lambda^3+\cdots}{(1-\lambda\gamma)} = \frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1-\lambda^3)}.
\end{aligned}$$

Notice that this γ-depreciated sum is equal to the λ-discounted sum when immediate rewards are scaled by a factor $\frac{1}{1-\lambda\gamma}$. We show that this is not a mere coincidence, and prove that this equality holds also in the context of general MDPs.

Theorem 3.1. *Over any finite MDP with λ-discounted value V_λ and λ-discounted γ-depreciating value V^γ_λ, it holds that $V^\gamma_\lambda = \frac{V_\lambda}{1-\gamma\lambda}$.*

The proof[1] of Theorem 3.1 decomposes the defining infinite series for V^γ_λ into a Cauchy product and then uses Mertens' Theorem (c.f. Theorem 3.50 of Rudin [16]) from the field of real analysis to establish convergence and characterize the limit. The following corollary states some direct consequences of Theorem 3.1 when combined with classic results from the literature [7, 9, 15, 18, 20].

Corollary 3.1. *For any discounted depreciating payoff with value V^γ_λ over a finite MDP M the following hold.*

(1) There exists a stationary deterministic optimal policy for V^γ_λ.
(2) Value V^γ_λ and optimal policies are computable in polynomial time.
(3) If each state-action pair in M is encountered infinitely often and learning rates satisfy the Robbins-Monroe convergence criteria, then Q-learning converges asymptotically to an optimal policy for V^γ_λ.

Thus, the reduction of discounted depreciating payoffs to standard discounted payoffs achieved by Theorem 3.1 provides the keystone for (i) proving the existence of simple optimal policies, (ii) establishing tractability of computing optimal values and policies, and (iii) obtaining a convergent reinforcement learning method with respect to discounted depreciating optimization.

4 AVERAGE DEPRECIATING PAYOFF

Next, consider the long-run average of the depreciating asset values as the limit inferior of the sequence

$$3,\ \frac{3\gamma+4}{2},\ \frac{3\gamma^2+4\gamma+5}{3},\ \frac{3\gamma^3+4\gamma^2+5\gamma+3}{4},\ \frac{3\gamma^4+4\gamma^3+5\gamma^2+3\gamma+4}{5},\ldots$$

Based on classical Tauberian results of Bewley and Kohlberg [3], it is tempting to conjecture that the λ-discounted, γ-depreciating value converges to this mean as $\lambda \to 1$ from below, i.e.

$$\begin{aligned}
\lim_{\lambda\to 1}(1-\lambda)\frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1-\lambda^3)} &= \lim_{\lambda\to 1}\frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1+\lambda+\lambda^2)} \\
&= \frac{3+4+5}{3(1-\gamma)}.
\end{aligned}$$

Indeed, we prove that this conjecture holds for general MDPs.

Theorem 4.1. *Over any finite MDP with long-run average γ-depreciating value V^γ and λ-discounted γ-depreciating value V^γ_λ, it holds that $V^\gamma = \lim_{\lambda\to 1}(1-\lambda)V^\gamma_\lambda$, where $\lambda \to 1$ from below.*

Since Mertens' Theorem fails to apply in this situation, the above result is obtained from a longer argument using only basic principles [6]. The next corollary collects and states a number of consequences of Theorem 4.1 (in combination with Theorem 3.1 and results from the aforementioned literature [7, 9, 15, 18]).

Corollary 4.1. *For any long-run average depreciating payoff with value V^γ over a finite MDP M the following hold.*

(1) There exists a stationary deterministic optimal policy for V^γ.
(2) Value V^γ and optimal policies are computable in polynomial time.
(3) If V is the long-run average value of M, then $V^\gamma = \frac{V}{1-\gamma}$.
(4) There exists a discount factor λ_0 such that, for any $\lambda \in [\lambda_0, 1)$, any policy optimal for V^γ_λ is also optimal for V^γ.

Hence, Theorem 4.1 allows us to lift the guarantees of items (1) and (2) of Corollary 3.1 from the setting of discounted depreciating optimization to that of long-run average depreciating optimization. Moreover, our result enables obtaining a remarkably simple characterization of the long-run average depreciating value in terms of the traditional long-run average value (item (3) of Corollary 4.1), mirroring the statement of Theorem 3.1. Lastly, the existence of Blackwell optimal policies (item (4) of Corollary 4.1) for long-run average depreciating optimization implies that Q-learning for discounted depreciating payoffs with sufficiently large discount factor λ also converges asymptotically to optimal policies for the long-run average depreciating payoff.

[1] See [6] for the full version, including precise theorem statements and complete proofs, of the present extended abstract.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. CCF-2146563.

REFERENCES

[1] Susan Rose Ackerman. 1973. Used cars as a depreciating asset. *Economic Inquiry* 11, 4 (1973), 463.
[2] Paul Beckerman. 1991. *The economics of high inflation.* Springer.
[3] Truman Bewley and Elon Kohlberg. 1976. The asymptotic theory of stochastic games. *Mathematics of Operations Research* 1, 3 (1976), 197–208.
[4] Oscar R Burt. 1972. A Unified Theory of Depreciation. *Journal of Accounting Research* (1972), 28–57.
[5] Pete Comley. 2015. *Inflation Matters: Inflationary Wave Theory, its impact on inflation past and present... and the deflation yet to come.* Pete Comley.
[6] Taylor Dohmen and Ashutosh Trivedi. 2023. Reinforcement Learning with Depreciating Assets. arXiv:2302.14176 [cs.AI] https://arxiv.org/abs/2302.14176
[7] Eugene A Feinberg and Adam Shwartz. 2012. *Handbook of Markov decision processes: methods and applications.* Vol. 40. Springer Science and Business Media.
[8] Adam Fergusson. 2010. *When Money Dies.* Old Street Publishing.
[9] Jerzy Filar and Koos Vrieze. 1996. *Competitive Markov decision processes.* Springer-Verlag.
[10] Shane Frederick, George Loewenstein, and Ted O'donoghue. 2002. Time discounting and time preference: A critical review. *Journal of economic literature* 40, 2 (2002), 351–401.
[11] Geoffrey Heal. 2007. Discounting: a review of the basic economics. *U. Chi. L. Rev.* 74 (2007), 59.
[12] Charles R Hulten and Frank C Wykoff. 1980. *The measurement of economic depreciation.* Urban Institute Washington.
[13] George Loewenstein and Elster Jon (Eds.). 1992. *Choice Over Time.* Russell Sage Foundation. 424 pages.
[14] Cédric Philibert. 1999. The economics of climate change and the theory of discounting. *Energy Policy* 27, 15 (1999), 913–927.
[15] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley. https://doi.org/10.1002/9780470316887
[16] Walter Rudin. 1976. *Principles of mathematical analysis.* McGraw-hill New York.
[17] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.
[18] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction.* MIT Press. https://www.worldcat.org/oclc/37293240
[19] R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). MIT Press.
[20] Christopher J. C. H. Watkins and Peter Dayan. 1992. Technical Note Q-Learning. *Mach. Learn.* 8 (1992), 279–292. https://doi.org/10.1007/BF00992698
[21] F Kenneth Wright. 1964. Towards a general theory of depreciation. *Journal of accounting research* (1964), 80–90.
[22] Frank C. Wykoff. 1970. Capital Depreciation in the Postwar Period: Automobiles. *The Review of Economics and Statistics* 52, 2 (1970), 168–172.